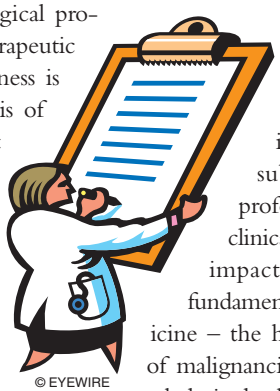
Application Notes

Paulo J.G. Lisboa, *Liverpool John Moores University, UK*
Alfredo Vellido, *Technical University of Catalonia, SPAIN*
Roberto Tagliaferri and Francesco Napolitano, *University of Salerno, ITALY*
Michele Ceccarelli, *University of Sannio, ITALY*
José D. Martín-Guerrero, *University of Valencia, SPAIN*
Elia Biganzoli, *University of Milano, ITALY*

Data Mining in Cancer Research

I. Introduction

Advances in cancer medicine have traditionally come from detailed understanding of biological processes, later translated into therapeutic interventions, whose effectiveness is established by rigorous analysis of clinical trials. Over the last two decades the increasing throughput of data from microarray screening, spectral imaging and longitudinal studies are turning the understanding of cancer pathology into as much a data-based as a biologically and clinically driven science, with potential to impact more strongly on evidence-based decision support moving towards personalized medicine [1].


© EYEWIRE

This article is not intended as a comprehensive survey of data mining applications in cancer. Rather, it provides starting points for further, more targeted, literature searches, by embarking on a guided tour of computational intelligence applications in cancer medicine, structured in increasing order of the physical scales of biological processes, as outlined in Table 1.

II. Robust Clustering and Data Visualization

Over the last decade, cancer has become a data-intensive area of research, with accelerating developments in data-acquisition technologies and specific methodologies, for example, *next-generation* sequencing for monitoring genetic changes in tumor cells as they progress from normal to invasive [2].

The first step in unravelling these processes is the study of co-occurrences, leading to the identification of disease sub-types. This already has profound implications for the clinical management of patients, impacting on one of the most fundamental precepts of cancer medicine – the histological characterisation of malignancies. Currently this is overwhelmingly done by measuring the differentiation between cancerous cells and the original normal cells from which they have evolved, marking them as well to poorly differentiated, meaning that the changes are linked with good to poor prognosis. Yet it is becoming apparent that the metabolism of cells is a key factor in uncontrolled cell proliferation, with a potentially greater influence on disease progression than cell differentiation alone.

Clustering is the first line analysis methodology to mine data bases for useful research hypotheses to take onto further, more directed, study [3], [4]. From a mathematical perspective the objects of study, be they genes, patients, mode of action drugs or disease sub-types, are points in a multi-dimensional space where similarity is defined, typically trough a measure of the distance between object pairs. The main classes of clustering approaches used in data mining for cancer are *partitional*, creating a simple partition of the data, or *hierarchical*, which create a tree-structured hierarchy of the data. Partitional approaches form a very

TABLE 1 Overview of healthcare interventions across physiological scales.

BIOLOGICAL PROCESSES	PHYSIOLOGICAL SCALE	INDUSTRIAL SECTOR	APPLICATION DOMAIN
MOLECULAR BIOLOGY	GENOME AND PHYSIOME	BIOINFORMATICS PHARMACEUTICS	SUSCEPTIBILITY TO DISEASE
METABOLIC PATHWAYS			TARGETS FOR INTERVENTION
CYTOLOGY AND HISTOLOGY	CELL AND TISSUE LEVELS	SYSTEMS BIOLOGY LABORATORY TESTS	DISEASE DIAGNOSIS
MEDICAL IMAGING	ORGAN LEVEL	NEUROINFORMATICS	ANATOMICAL AND PHYSIOLOGICAL MONITORING
ELECTROPHYSIOLOGICAL MEASUREMENT		MEDICAL INFORMATICS	
CLINICAL SIGNS	SYSTEM LEVEL	MEDICAL EQUIPMENT & INSTRUMENTATION	DIAGNOSIS, PROGNOSIS AND SCREENING
OUT-OF-HOSPITAL CARE	LEVEL OF THE INDIVIDUAL	PHARMACEUTICS	AMBULATORY MONITORING FOR EARLY DIAGNOSIS
LIFESTYLE FACTORS	POPULATION LEVEL	PUBLIC HEALTH INFORMATICS	DISEASE PROTECTION AND PREVENTION

Digital Object Identifier 10.1109/MCI.2009.935311

1556-603X/10/$26.00©2010IEEE

heterogeneous class, including techniques based on Global Optimization, Vector Quantization, Kernel functions, Fuzzy sets, and Graph theory [5]. Hierarchical approaches can be *agglomerative* or *divisive* and are often used as for first-line clustering of high-throughput data. However, high-dimensional data spaces are by nature ultra-metric, each data point appearing to be at the centre of a hollow sphere, in contrast with the usual intuitive picture in low dimensions [6]. This effect needs to be taken into account especially with agglomerative methods to avoid early stage errors that can arise, resulting in sub-optimal solutions compared with partition algorithms. A third main class of algorithms is biclustering that is to say clustering of both objects and features within a single model [7]. The range of clustering techniques illustrates the complexity of the problem.

The fundamental concept underlying the definition of clusters, "similarity", is complex in itself because of its ambiguity. This difficulty is compounded when the most relevant features are not known *a priori,* yet the choice of different features usually produces different solutions. An open research question is *feature selection* operating with unsupervised training. This will normally involve parametric modeling and the application of Bayesian methods to derive the discriminative features which best separate between the clusters [8].

In addition to such *intrinsic* complexity, purely algorithmic issues may give rise to different solutions for the same data set, even for the same algorithm. This raises the possibility that cluster stabilization methods including *global stability* assessments [10], [11] and *consensus clustering* [12], [13] methods may miss the real complexity of the problem.

An alternative approach to clustering assessment shifts the focus from the search of the best solution to the search of a set of equally plausible, though different, solutions [14]. Recently, meta-clustering [15] (see fig 1) has been proposed as a clustering assessment tool. It is the process of clustering the solutions i.e., the cluster partitions themselves, generalizing the concept of global stability to *local* stability (see Fig. 1), that is, the stability of a solution with respect to a homogenous subset of the candidates.

The cluster composition then needs to be explored in detail, for which dendrograms provide a useful representation, usually through sequential univariate splitting of the data space. However, the interpretation of geometrical relationships in the data is often best derived from direct spatial representations. The most commonly used methods for *data visualization* are Multi Dimensional Scaling, Principal Component Analysis (PCA) and Self-Organizing Maps [16], [17].

Linear methods are particularly interesting since they enable axes projections to be overlaid on the data. An efficient methodology for cluster-based linear visualization is to use a decomposition of the covariance matrix which explicitly takes into account cluster membership. This is the natural tailoring of PCA to the case where the data are partitioned into labelled cohorts – an example from breast cancer proteomics is shown in Fig. 2 [18].

III. Pathway Modeling

The study of the cell functions is modelled from experimental data sets which are both complex and dynamical [19], [20]. This has spawned the field of *Systems Biology* to understand and describe how molecular components interact to manifest the phenotypic behaviour which is the target of disease intervention [21], illustrated in Fig. 3. In the systems biology literature there are two main classes of approaches:

- ❏ Inverse problems, which involves mining data from many perturbation experiments to identify a compatible structure of the signaling pathway.
- ❏ Direct problems, to forecast and simulate the forward evolution of dynamical systems, using differential equations.

Several methods make the assumption that the expression level of a given gene can be considered as a random variable and the mutual relationships between them can be derived by statistical dependencies. The resulting system is embedded in a probabilistic Graphical Model [22] described for example by a Bayesian Network [23], a Factor Graph [24], a Gaussian Graphical Model [25] or a Mutual Information derived Influence Network [26].

IV. Cancer Imaging

Cancer imaging is a vast field focused on non-invasive measurement for

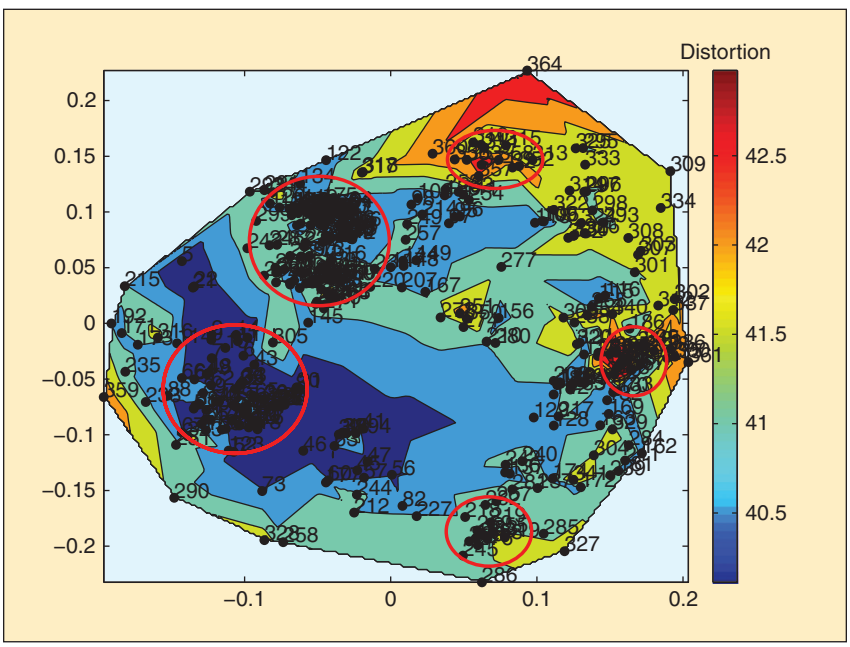


FIGURE 1 An illustration of meta-clustering. Each point on the map represents a clustering solution and close points represent similar solutions. Colors indicate a measure of quality for the solutions and red circles highlight local stability [15].

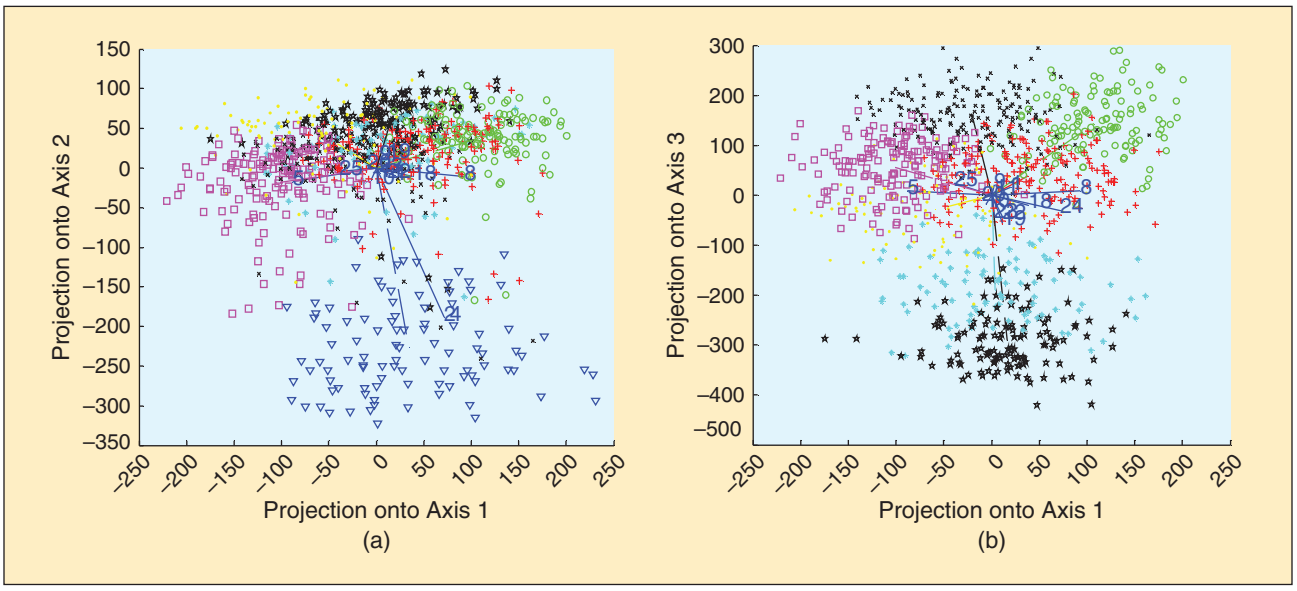


FIGURE 2 Two views of cluster-based visualization of 25-d protein expression date from breast cancer histology showing (a) the HER-2 positive cohort distinct from the rest and (b) rotating the 3-d view to illustrate the separation between the projections of the remaining seven clusters [18].

cancer detection, diagnosis and grading, and to monitor response to therapy by measuring tumor size and detecting regrowth. Having started out as mainly anatomical measurement for anomaly detection as an indicator of malignancy, for instance using ultrasound and Magnetic Resonance Imaging (MRI), cancer imaging has developed into multiple modalities of physiological imaging including spectroscopy, with Magnetic Resonance Spectroscopy (MRS) and spatially localized Chemical Shift Images (CSI), with Positron-Emission Tomography (PET) and Single-Photon emission CT (SPECT) clinically widely used to measure glucose metabolism and therefore find regions with exceptionally high energetic rates, which are characteristic of tumor growth. An accessible review of machine learning applications to detection and diagnosis of disease may be found in [27].

A recent development is the use of specific functional imaging for tumor delineation, that is to say, to measure the contour during surgical planning and to identify tumor striations that may not be apparent clinically even during re-section. Initial efforts in this direction applied blind signal separation to single-voxel spectra to separate infiltrating tumor mass compared with necrotic tissue [28]. However, tumors are not homogeneous, often mixing different types of tissue and sometimes with different grades of malignancy. One way to resolve this mixture is to segment the tumor composition, which is called nosological imaging, of which examples of diagnostic applications in brain are reported in [29], [30]. A related and very important clinical area is follow-up imaging to determine tumor progression and to differentiate recurrent tumor growth from treatment-induced tissue changes [31], [32].

V. Prognostic Outcome

Outcome modeling usually refers to the analysis of patterns of mortality and recurrence in longitudinal cohort studies, where a specific groups of patients followed-up for up to 20 years, recording the occurrence of events of interest which may be cancer specific mortality, overall mortality or local and distal recurrence. The value of these studies is closely linked to the design of the follow-up process, highlighting the need to involve biostatistical expertise from the very earliest stages in study design.

It is required to model the likelihood of the event of interest, conditional on the event not taking place at the start of the time interval and taking into account the loss of patients to follow-up for reasons other than the event of interest, termed censorship.

The application in medical statistics of linear-in-the-parameters models to non-linear effects in outcome data generally involves discretising the data space to generate piecewise linear approximations to the desired response surfaces. However, this heuristic approach can have important consequences in cancer research, making it difficult to mine the hazard function for insights into the dynamics of metastatic spread [33] and inducing a categorisation of the histological grade of the tumor, which can give rise to important subjective effects that impact on the choice of therapy given to the patient [34].

Neural network models fit non-linear, time dependent estimates of the hazard function, without the need to make proportionality assumptions or to discretize the state-space [35], [36] and their accuracy has been established by multi-centre, double-blind, evaluation [37] and this framework has recently been extended for estimating the joint model for more than one risk, e.g. local and distal recurrence [38].

VI. Conclusion

Computational intelligence methodology has been widely applied to data

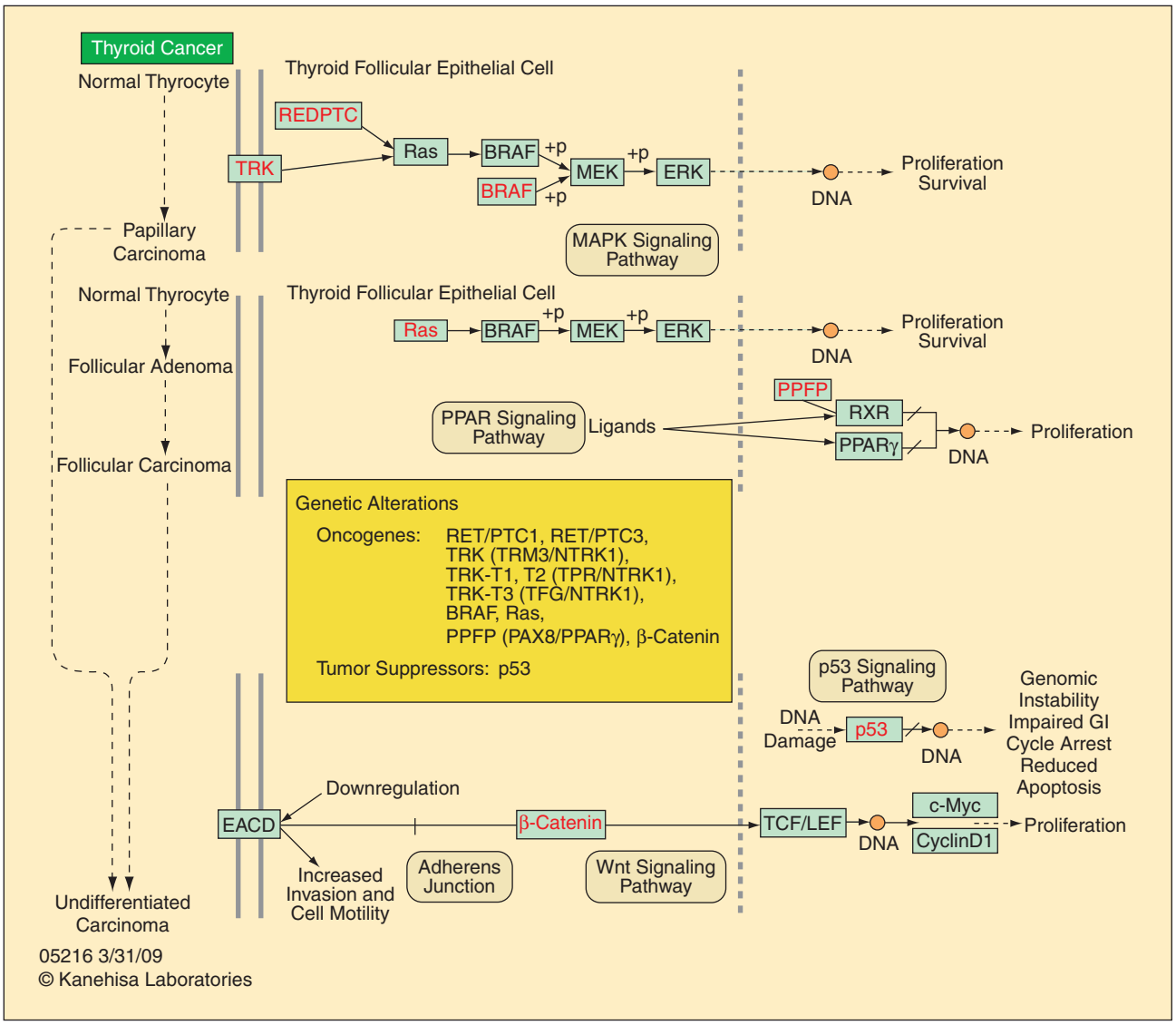


FIGURE 3 An excerpt of the papillary thyroid carcinoma pathway from the KEGG database.

mining in cancer but open questions remain before we can fully exploit the potential of molecular biology, to integrate this into routine imaging and other non-invasive measurements for screening and early diagnosis, also to deliver consistency in laboratory measurements especially in histopathology, and to more accurately model the balance between benefit and risk of therapy for niche groups of patients, down to the level of the individual.

New data mining methods need to be developed to handle high-dimensionality and large data volumes, to more efficiently model the dynamics of deep molecular pathways with high levels of noise and small experimental sample sizes, and also to robustly estimate hazard functions with censored data with multiple competing risks. Analytical models need also to be integrated with clinical expert knowledge and processes, potentially by mapping into Boolean filters [39].

The road ahead is long and bumpy. This article is provided as a sign-post to the nature of the key methodological and application challenges in this growing and exciting field.

Acknowledgments

This work was supported in part by the Biopattern NoE under FP6/2002/IST/1 and IST-2002-508803. Further support was provided by MICINN research project TIN2009-13895-C02-01.

References

[1] P. J. G. Lisboa, A. F. G. Taktak, "The use of artificial neural networks in decision support in cancer: A systematic review," *Neural Netw.*, vol. 19, pp. 408–415, 2006.
[2] P. J. Campbell, et al., "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nature Genetics*, vol. 40, pp. 722–729, 2008.
[3] N. Belacel, Q. C. Wang, and M. Cuperlovic-Culf, "Clustering methods for microarray gene expression data," *Omics*, vol. 10, no. 4, pp. 507–531, 2006.
[4] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Trans. Knowledge Data Eng.*, vol. 16, no. 11, pp. 1370–1386, 2004.
[5] D. Wunsch and R. Xu, *Clustering (IEEE Press Series on Computational Intelligence)*. Washington, DC: IEEE Computer Society Press, 2008.
[6] F. Murtagh, G. Downs, and P. Contreras, "Hierarchical clustering of massive, high dimensional data sets by

exploiting ultrametric embedding," *SIAM J. Sci. Comput.*, vol. 30, pp. 707–730, 2008.
[7] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Trans. Computat. Biol. Bioinform.*, vol. 1, no. 1, pp. 24–45, 2004.
[8] L. Carrivick, et al., Identification of prognostic signatures in breast cancer microarray data using Bayesian techniques," *J. R. Soc. Interface*, vol. 3, no. 8, pp. 367–381, 2006.
[9]. A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Biocomputing (Proc. Pacific Symp.)*, vol. 7, R. B. Altman and K. Lauderdalc, Eds. Kauai, Hawaii, USA, 2002, pp. 6–17.
[10] M. K. Kerr and G. A. Churchill, "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments," *Proc. Natl. Acad. Sci.*, vol. 98, pp. 8961–8965, 2001.
[11] L. I. Kuncheva and D. P. Vetrov, "Evaluation of stability of k-means cluster ensembles with respect to random initialization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 11, pp. 1798–1808, 2006.
[12] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, 2007.
[13] N. Nguyen and R. Caruana, "Consensus clustering," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, 2007, pp. 607–612.
[14] R. Caruana, M. Elhawary, N. Nguyen, and C. Smith, "Meta clustering," in *Proc. 6th IEEE Int. Conf. Data Mining (ICDM)*, 2006, pp. 107–118.
[15] A. Ciaramella, et al., "Interactive data analysis and clustering of genomic data," *Neural Netw.*, vol. 21, pp. 368–378, 2008.
[16] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. New York: Springer-Verlag, 2007.
[17] J. Vesanto, "SOM-based data visualization methods," *Intell. Data Anal.*, vol. 3, pp. 111–126, 1999.
[18] P. J. G. Lisboa, I. O. Ellis, A. R. Green, F. Ambrogi, and M. B. Dias, "Cluster-based visualisation with scatter matrices," *Pattern Recogn. Lett.*, vol. 29, no. 13, pp. 1814–1823, 2008.
[19] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins, "Computational studies of gene regulatory networks: In numero molecular biology," *Nature Rev. Gene.*, vol. 2, pp. 268–279, 2001.
[20] H. de Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *J. Computat. Biol.*, vol. 9, no. 1, pp. 67–103, 2002.
[21] E. J. Borowski and J. M. Borwein, *Computational Modeling of Genetic and Biochemical Networks*. Cambridge, MA: MIT Press, 2001.
[22] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, p. 799, 2004.
[23] C. Needham, J. Bradford, A. Bulpitt, and D. Westhead, "A primer on learning in Bayesian networks for computational biology," *PLOS Computat. Biol.*, vol. 3, no. 8, pp. 1409–1416, 2007.
[24] I. Gat-Viks, A. Tanay, D. Raijaman, and R. Shamir, "A probabilistic methodology for integrating knowledge and experiments on biological networks," *J. Computat. Biol.*, vol. 13, no. 2, pp. 165–181, 2006.
[25] J. Schafer and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics*, vol. 21, no. 6, pp. 754–764, 2005.
[26] K. Basso, et al., "Reverse engineering of regulatory networks in human B cells," *Nature Gene.*, vol. 37, no. 4, pp. 382–290, 2005.
[27] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006.
[28] Y. Huang, P. J. G. Lisboa, and W. El-Deredy, "Tumour grading from magnetic resonance spectroscopy: A comparison of feature extraction with variable selection," *Stat. Med.*, vol. 22, no. 1, pp. 147–164, 2003.
[29] F. Szabo De Edelenyi, et al., "A new approach for analyzing proton magnetic resonance spectroscopic images of brain tumors: Nosologic images," *Nature Med.*, vol. 6, pp. 1287–1289, 2000.
[30] M. De Vos, T. Laudadio, A. W. Simonetti, A. Heerschap, and S. Van Huffel, "Fast nosologic imaging of the brain," *J. Magnet. Reson.*, vol. 184, no. 2, pp. 292–301, 2007.
[31] A. H. Jacobs, et al., "Imaging in neurooncology," *J. Amer. Soc. Exp. NeuroTherapeu.*, vol. 2, pp. 333–347, 2005.
[32] S. Cha, "Neuroimaging in neuro-oncology," *J. Amer. Soc. Exp. NeuroTherapeu.*, vol. 6, pp. 465–477, 2009.
[33] R. Demicheli, P. Valagussa, and G. Bonadonna, "Double-peaked time distribution of mortality for breast cancer patients undergoing mastectomy," *Breast Cancer Res. Treat.*, vol. 75, pp. 127–134, 2002.
[34] J. A. Woolgar, "The predictive value of detailed histological staging of surgical resection specimens in oral cancer," in *Outcome Prediction in Cancer*, A. F. G. Taktak and A. C. Fisher, Eds. U.K.: Elsevier, 2007, pp. 3–26.
[35] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach," *Stat. Med.*, vol. 17, pp. 1269–1206, 1998.
[36] A. Eleuteri, R. Tagliaferri, L. Milano, S. De Placido, and M. De Laurentiis, "A novel neural network-based survival analysis model," *Neural Netw.*, vol. 16, no. 5–6, pp. 855–864, 2003.
[37] A. Taktak, et al., "Double-blind evaluation and benchmarking of survival models in a multi-centre study," *Comput. Biol. Med.*, vol. 37, no. 8, pp. 1108–1120, 2007.
[38] P. J. G. Lisboa, et al., "Partial logistic artificial neural network for competing risks regularised with automatic relevance determination," *IEEE Trans. Neural Netw.*, vol. 20, no. 9, pp. 1403–1416, 2009.
[39] T. Rögnvaldsson, T. A. Etchells, L. You, D. Garwicz, I. H. Jarman, and P. J. G. Lisboa, "How to find simple and accurate rules for viral protease cleavage specificities," *BMC Bioinform.*, vol. 10, pp. 149, 2009.